Article

Evaluation Review
2016, Vol. 40(4) 279-313

DOI: 10.1177/0193841X16655657
journals.sagepub.com/home/erx
SAGE

Academic and Behavioral Design Parameters for Cluster Randomized Trials in Kindergarten: An Analysis of the Early Childhood Longitudinal Study 2011 Kindergarten Cohort (ECLS-K 2011)

E. C. Hedberg[1]

Abstract

Background: There is an increased focus on randomized trials for proximal behavioral outcomes in early childhood research. However, planning sample sizes for such designs requires extant information on the size of effect, variance decomposition, and effectiveness of covariates. **Objectives:** The purpose of this article is to employ a recent large representative sample of early childhood longitudinal study kindergartners to estimate design parameters for use in planning cluster randomized trials. A secondary objective is to compare the results of math and reading with the previous kindergartner cohort of 1999. **Research Design:** For each measure, fall–spring gains in effect size units are calculated. In addition,

[1] NORC, University of Chicago, Chicago IL, USA

Corresponding Author:
E. C. Hedberg, NORC, University of Chicago, 1155 E 60th Street, Chicago, IL 60637, USA.
Email: hedberg-eric@norc.org

multilevel models are fit to estimate variance components that are used to calculate intraclass correlations (ICCs) and R^2 statistics. The implications of the reported parameters are summarized in tables of required school sample sizes to detect small effects. **Measures:** The outcomes include information about student scores regarding learning behaviors, general behaviors, and academic abilities. **Results:** Aside from math and reading, there were small gains in these measures from fall to spring, leading to effect sizes between about .1 and .2. In addition, the nonacademic ICCs are smaller than the academic ICCs but are still nontrivial. Use of a pretest covariate is generally effective in reducing the required sample size in power analyses. The ICCs for math and reading are smaller for the current sample compared with the 1999 sample.

Keywords
education, methodological development

Introduction

With the increased focus in education research on evidence derived from randomized trials, the cluster randomized trial is now the standard for most evaluations. A cluster randomized trial is where entire clusters, for example, schools or classrooms, are assigned to treatment conditions where the outcomes are measured at the unit level, for example, students. Success in rejecting the null hypothesis is quantified as statistical power (Cohen, 1992), which measures the probability of detecting a true effect (1—Type II error) given a sampling design (e.g., number of clusters and units per cluster) and desired Type I error (α). Planning cluster randomized trials requires extant information to estimate the appropriate sample size and power, including design parameters of the type detailed in this article. In the past 10 years, there has been an increase in the amount of work outlining design parameters.[1]

In early childhood, research has clearly demonstrated that nonacademic outcomes such as internal/externalizing behaviors and early cognition lead to later academic success (see, e.g., Claessens, Duncan, & Engel, 2009; Denham, 2006; Denham & Brown, 2010; DiPerna, Lei, & Reid, 2007). As a result, several interventions and programs are targeting not only the distal outcome of academic achievement but also these proximal behavioral outcomes. For example, the Corporation for National Community Service has

several national and state-specific programs that request proposals targeting, in part, social and behavioral outcomes (e.g., http://www.in.gov/serveindiana/files/IN_-_2016_ASN_Notice.pdf). The Institute for Education Sciences also has a long-term grant program that supports research on special education and behavioral outcomes (see https://ies.ed.gov/funding/ncser_progs.asp).

These interventions target nonacademic outcomes in early childhood as a method to produce higher academic gains later in life. However, interim evaluations of these interventions need to test whether differences in the nonacademic outcomes have been achieved, and so these studies need to be powered to detect these effects. Thus, the focus of this article is on kindergartners, because even though nonacademic outcomes are more proximal, the act of randomizing entire clusters (schools) to treatment or control still means that the analysis is bound by the parameters of a multilevel design. As such, planning studies requires the same set of expectations about variance decomposition, effectiveness of covariates, and effect sizes that studies targeting academic outcomes require.

Collection of evidence about nonacademic outcomes in early childhood is ongoing. Some evaluations report parameters for social emotional outcomes (e.g., Rhoades, Greenberg, & Domitrovich, 2009), and systematic collections of parameters for these outcomes, such as Jacob, Zhu, and Bloom (2010), also offer important guides. Unfortunately, the estimates that have been published are the result of limited samples from experiments, and so analysis of a large representative sample, analogous to Hedges and Hedberg (2007), is warranted.

Many researchers suppose that the clustering of proximal outcomes is similar to the clustering behaviors of typical distal outcomes such as math and reading scores. The What Works Clearinghouse (WWC; 2014), for example, points reviewers to an estimate of 10% for the total variation in nonacademic outcomes at the school level when making cluster-based adjustments to statistical tests. These adjustments are required for evaluating research that improperly tests hypotheses without taking the clustered nature of the data into account (Hedges, 2007a). Typically, the WWC proposes 20% for academic outcomes and 10% for nonacademic outcomes. The guidance for academic outcomes is well grounded in empirical evidence (e.g., Hedges & Hedberg, 2007), but the guidance for nonacademic outcomes is less grounded. The guidance 10% for nonacademic outcomes may be too high.

While clustering of academic measures such as math and reading is well documented (e.g., Hedges & Hedberg, 2007), it is plausible that behavioral outcomes are less dependent on organizational efforts and thus will exhibit smaller school effects. In other words, since schools' main goal is to work toward improving academic outcomes, behavioral outcomes of young children may be less impacted by school activities. Thus, the between-school variance of outcomes such as internalizing or externalizing behaviors may be small and so the guidance suggesting that 10% of the variation in between schools may be too high. On the other hand, the practical experience of early childhood education involves many nonacademic activities specifically designed to improve nonacademic outcomes, and some activities may impact more proximal, behavioral outcomes. What is needed then are estimates of these crucial design parameters to plan cluster randomized trials seeking to evaluate proximal nonacademic outcomes.

In addition to parameters about variance decomposition, researchers require knowledge about typical growth in the scale of standardized difference in means effect sizes. Bloom, Hill, Black, and Lipsey (2008) have provided empirical benchmarks for a year's worth of growth in effect size units for math and reading. Their results indicate that reading ability grows by 1.5 standard deviations (*SDs*) while math growth slightly greater than 1 *SD*. Again, little is systematically known about nonacademic outcome gains from year to year.

The Present Study

The purpose of this article is to employ the early childhood longitudinal study, kindergarten class of 2010–2011 (ECLS-K: 2011, see Mulligan, Hastedt, & McCarroll, 2012, for an overview), to estimate design parameters useful in planning cluster randomized trials.[2] The outcomes include information about student scores regarding learning behaviors (approaches to learning and attentional focus), general behaviors (externalizing problem behaviors, inhibitory control, internalizing problem behaviors, interpersonal skills, and self-control), cognitive abilities (card sort postswitch score, card sort border score, and numbers reversed score), and academic abilities (math and reading scores). The parameters are presented for the spring scores (after a year's worth of instruction by the school).

The present article, then, has several objectives. First, it will outline the gains in the outcome measures in effect size units for use as upper bound reference points. Second, it will present intraclass correlations (ICCs) and R^2 statistics for each outcome's spring score. Next, to contextualize these

results, estimated school-level sample size requirements are then calculated for each outcome and set of covariates for the effect sizes that reflect the analysis of the nonacademic outcomes. The goal of the sample size table is not to provide guidance on exact sample sizes per se but instead to offer guidance on the feasibility of conducting studies for the given effect sizes and design parameters. Actual planned sample sizes will be different depending on the number of students and the expected effect size. In addition, this article updates the Hedges and Hedberg (2007) estimates for math and reading using the new 2011 ECLS-K cohort and explores the changes to these parameters between cohorts.

This article is organized as follows. First, the sample and outcomes are reviewed. Next, the statistical methodology and analysis choices are covered. This article then presents the results and then ends with discussion. A review of the parameters necessary for planning the appropriate analysis of a cluster randomized trial is included in Appendix A.

Study Sample and Outcomes

The ECLS is a series of studies carried out by the U.S. Department of Education to better understand the dynamics of early childhood. The original kindergarten cohort was examined during the 1998–1999 school year. The present cohort was examined during the 2010–2011 school year. These data provide a rich pool of information from students, teachers, parents, and caregivers on a variety of topics. Of interest to this study are the teacher reports of several classroom behaviors (and corresponding parent reports when available) and the standardized tests (Tourangeau et al., 2012). The data set released contains 18,174 student records and 1,328 schools recorded for the spring scores. For this study, all students with spring test scores were selected for each outcome (i.e., the N is different for different test scores); 10,653 (59%) student records contained all 14 outcomes and 25% were missing one or two scores; and 523 students (3%) were missing all 14 spring outcomes. Appendix Table B1 gives the proportion of students missing test scores for each outcome and by the sample characteristics described below. Broadly, minority students were more likely to have missing data on all outcomes, as were students in the northeast, west, and urban areas.

Table 1 presents an overview of each measure, the developers, and the spring score reliability measure. Each measure is organized so that larger scores indicate positive progress, and thus positive effect sizes indicate gains in the desired direction. The outcomes in this study are organized

Table 1. Description of Analyzed Outcomes.

Outcome	Developers	Reliability	Description
Learning behaviors			
Approaches to learning (TR)	ECLS-K survey team	.91	Scores indicate whether the child is more organized, eager to learn new things, works independently, adapts to changes, persists in completing tasks, pays attention, and follows classroom rules.
Approaches to learning (PR)	ECLS-K survey team	.72	Scores indicate whether the child is able to persist and complete various tasks.
Attentional focus	Rothbart et al. (2001); Putnam and Rothbart (2006)	.87	Higher scores indicate child is able to focus attention.
General behaviors			
Externalizing problem behaviors	Gresham and Elliott (1990)	.89	Measures disturbing activities and fighting with others. Higher scores mean a higher frequency of problem behaviors.
Inhibitory control	Putnam and Rothbart (2006; Rothbart et al. (2001)	.87	Scores indicate whether the child was able to resist the inclination to do something inappropriate.
Internalizing problem behaviors	Gresham and Elliott (1990)	.78	Scores indicate frequent problems such as low self-esteem and depressed behavior.
Interpersonal skills	Gresham and Elliott (1990)	.87	Scores indicate interpersonal skills of the child.
Self-control (TR and PR)	Gresham and Elliott (1990)	.82	Measures persisting in activities and following classroom rules.
Cognitive abilities			
Card sort postswitch score	Zelazo (2006)	NA	Children are asked to sort cards based on color. The summary score indicates higher cognitive functioning.

(continued)

Table 1. (continued)

Outcome	Developers	Reliability	Description
Card sort border score	Zelazo (2006)	NA	Children are asked to sort cards based on color, shape, and border. The summary score indicates higher cognitive functioning.
Numbers reversed score	Woodcock (1990)	NA	Measures the students working memory through a backward digit span.
Academic abilities			
Math IRT score	Mulligan, Hastedt, and McCarroll (2012)	NA	Measures aspects of conceptual knowledge, procedural knowledge, and problem-solving.
Reading IRT score	Mulligan et al. (2012)	NA	Measures disturbing activities and fighting with others. Higher scores mean a higher frequency of problem behaviors.

Note. ECLS-K = early childhood longitudinal study kindergartners; IRT = item response theory; NA = not applicable; PR = parent report; TR = teacher report.

around four types: learning behaviors, general behaviors, cognitive abilities, and academic abilities. Learning behaviors include approaches to learning, developed by the ECLS survey team and Rothbart, Ahadi, Hershey, and Fisher's (2001) attentional focus measure. Next, general behaviors is a set of measures regarding both internalizing and externalizing problem behaviors (Gresham & Elliott, 1990), Rothbart, Ahadi, Hershey, and Fisher's (2001) inhibitory control, interpersonal skills, and self-control (Gresham & Elliott, 1990). Cognitive ability measures include Zelazo's card sort games (2006) and the numbers reversed measure developed by Wookcock in 1990. Finally, academic outcomes are comprised of math and reading item response theory (IRT) scores.

Many of these outcomes are the focus of interventions. For example, Stormont (2002) outlines several avenues for interventions to improve externalizing problem behaviors at the teacher and classroom levels. Barrera et al. (2002) implemented a randomized trial of an intervention designed to reduce both internalizing and externalizing problem behaviors.

As a cognitive outcome example, Smith et al. (2013) used the numbers reversed metric, in part, to assess whether physical activity alleviated issues related to attention deficient disorders.

Statistical Method

In this section, the methods and formulas that are employed to estimate the design parameters are discussed. For each outcome, the same methods are used in the estimation process, which employed Stata 14's ''mixed'' restricted maximum likelihood (REML) estimation procedures. Each estimate is associated with an estimate of the sampling variance. For all the estimates, weights were not employed. This is analogous to the procedures in Hedges and Hedberg (2007) and the reasons are discussed at the end of this section. Other topics in this section include issues of assumptions and reproducibility.

Effect Size Benchmarks

Estimates of the effect size for fall–spring gains employed the difference between the spring and fall mean, $\bar{y}_{\text{spring}} - \bar{y}_{\text{fall}}$, standardized by the spring *SD*, s_{spring},[3]

$$\text{ES} = \frac{\bar{y}_{\text{spring}} - \bar{y}_{\text{fall}}}{s_{\text{spring}}}.$$

ICCs

ICCs are estimated for the spring scores. To estimate the ICC for a given outcome, *y*, a multilevel model is fit for the *i*th observation in the *j*th school

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij},$$

where the software computes the REML estimates of the variance of u_{0j}, which is $\hat{\sigma}_2^2$, and the variance of e_{ij}, which is $\hat{\sigma}_1^2$.

The estimate of the ICC, ρ, is then[4]

$$\hat{\rho} = \frac{\hat{\sigma}_2^2}{\hat{\sigma}_2^2 + \hat{\sigma}_1^2}.$$

R^2 Estimates

R^2 estimates for three sets of covariates are estimated. The first set of covariates is the fall score and its school mean, the second set of covariates are demographic indicators for gender, race, socioeconomic status, and their school means. The final set of covariates combines the fall score, demographic indicators, and the associated school means.

In order to estimate the R^2 statistics for each set of covariates, a model is fit to the data where the spring score for the ith student in school j is

$$y_{ij} = \gamma_{00} + \sum_p \gamma_{p0} x_{pij} + \sum_p \gamma_{0p} \bar{x}_{pj} + u^*_{0j} + e^*_{ij},$$

where the software computes the REML estimates of the variance of u^*_{0j}, which is $\hat{\tilde{\sigma}}^2_2$, and the variance of e^*_{ij}, which is $\hat{\tilde{\sigma}}^2_1$. The estimates of the R^2 statistics are then

$$\hat{R}^2_2 = 1 - \frac{\hat{\tilde{\sigma}}^2_2}{\hat{\tilde{\sigma}}^2_2},$$

at the school level and

$$\hat{R}^2_1 = 1 - \frac{\hat{\tilde{\sigma}}^2_1}{\hat{\tilde{\sigma}}^2_1},$$

at the student level.[5]

Estimates of Required Schools

To illustrate the implications of the estimated design parameters, the number of schools required to detect an effect size of .1 or .2 is estimated. This is the result of an iterative process, where for each outcome the appropriate design parameters were entered into the noncentrality formula reviewed in the appendix, the number of students was set to 15 and 30, and the number of schools was incremented by 2 (one for treatment and one for control) until the exact power equaled or exceeded .8 with a Type I error rate of .05 for a two-tailed test.

Mixed Model Assumptions

Mixed models make the assumption that the variances of the random effects, u_{0j} and e_{ij}, are normally distributed. Thus, in order to validate the results of the ICC analysis, the distributions of the student residuals and

school means were examined visually. The values that are analogous to e_{ij}, the student residuals, were calculated by simply subtracting the school mean of each score value from the student-level score. The distribution of u_{0j} was checked by examining the distributions of the school means. These distributions were checked using kernel density plots for each outcome and are presented in Figure 1. While all outcomes, student residuals, and school means failed a formal Shapiro-Wilk test of normality (1965; z scores of the tests are reported in Figure 1), most outcomes show approximate normal distributions for student residuals and school means, with the exception of the card sort postswitch score, in which the school means follow the overall distribution where most students score highly.

Reproducibility

As with all research, others may wish to replicate the results of the current article. This is encouraged. However, it should be noted that the variance components will be different with the longitudinal data sets (compared to the base-year only files) because the earlier observations (such as kindergarten) are rescaled to vertically equate with the later grades (such as first grade). This is broadly described on page 5-1 of the eighth grade psychometric report for the first ECLS-K cohort (Najarian, Pollack, Sorongon, & Hausken, 2009).[6] Given this issue of reproducibility, it is noted that the results of this article are based on the restricted use base-year file and not the publicly released K-1 file, which will produce a slightly different set of answers.

The Decision to Not Use Sampling Weights to Estimate Design Parameters

The ECLS-K 2011 sample of schools is not a simple random sample of clusters. The data collection team first selected counties (or county groups) as the primary sampling units (PSUs) using a combination of sampling proportionate to size (to ensure representation of the largest PSUs) and stratified sampling to ensure representation of the variety of school settings (Tourangeau et al., 2012). As a result, several types of settings are overrepresented or underrepresented based on the sampling parameters (detailed in the data documentation). Thus, these data do not represent a simple random sample of schools. To estimate design parameters, the issue is not as simple as just using design weights, however. The scaling decisions about how to handle the Level 1 (Student) weights (which are not designed for multilevel modeling because they include the combined sampling probabilities of the school

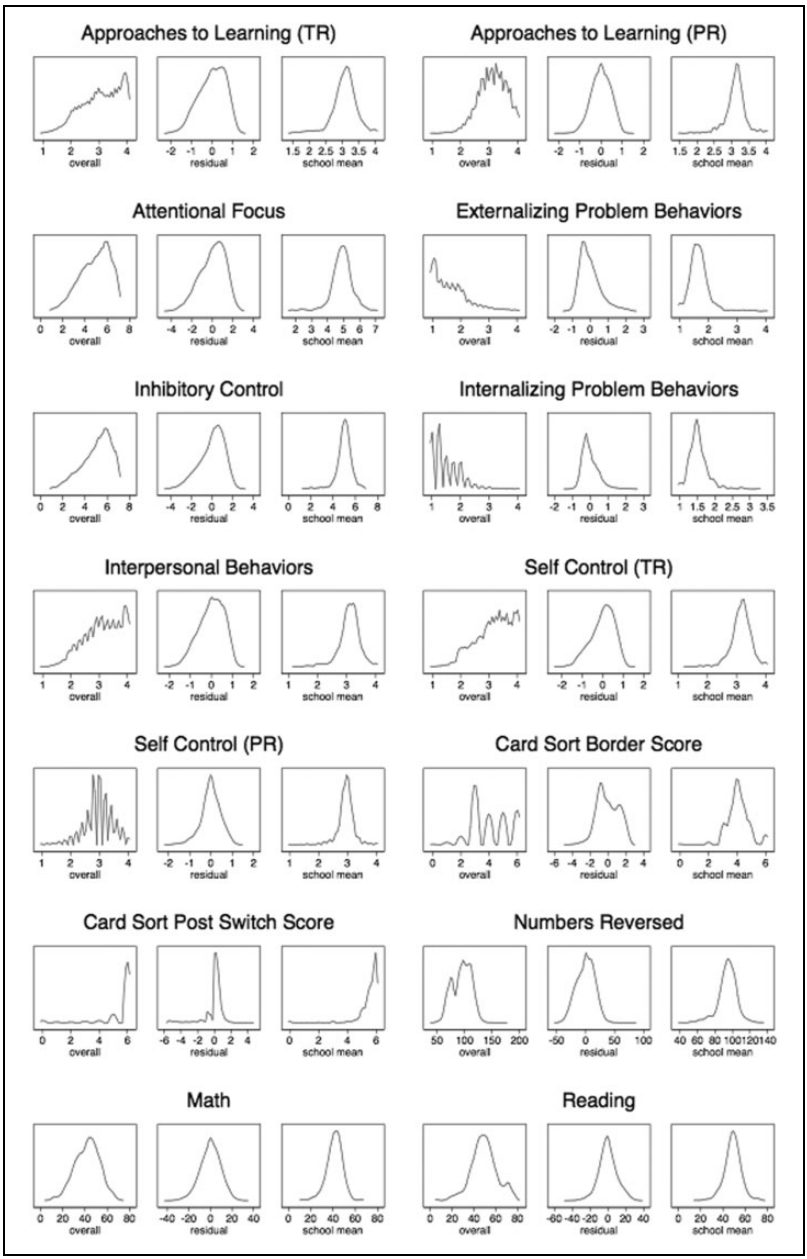

Figure 1. Distribution of spring scores, student residuals, and school means for each outcome.

selection and student selection given the school selection) can have large impacts on the design parameters. Furthermore, the use of school-level weights can bias the between-school variance components (Rabe-Hesketh & Skrondal, 2006). In the end, the literature offers no absolute advice as yet.

Regardless of the technical issues surrounding survey data weights and design parameters, the practical reality is that while the weights may be appropriate for the estimation of *fixed* effects (regression slopes, etc.), the supplied weights may or may not be appropriate for estimating variance components of *random* effects. Moreover, the *only* school-level weight available at this time is the school-base weight for the school administrator survey that is adjusted for nonresponse. Several schools that collected data on the outcomes of interest did not have corresponding administrator surveys. The result is that 462 of the 1,328 schools (35% of schools and 12% of students) in the study have a school-level weight of zero. An analysis of these schools indicated that while the average reading score, for example, was slightly higher than for schools with nonzero weights, the between-school variance for the schools with a zero weight is much larger than those with nonzero weights. Thus, the use of the nonresponse adjusted school-level administrator reduces the ICC estimate with less data. Moreover, since none of the provided weights are designed for use in estimating variance components and since the misuse of design weights has as-yet known impacts on estimated design parameters, the choice of this article is to simply not employ the weights in the analysis.

Results

Sample and Descriptive Measures

Table 2 presents the basic demographic characteristics of the employed sample (2011) and its comparison to the 1999 ECLS-K cohort. All characteristics are presented as the percentage of a categorical variable. The 2011 sample is 49% female as it was in 1998. The poverty level of the 2011 sample is higher than the 1999 sample, increasing from 19% to 26% below the poverty threshold. Age is consistent between samples, with most of the spring ages falling between 74 months and 78 months. The 2011 sample is more diverse: Whites are no longer the majority, and there is a marked increase in Hispanic student (18–25%). The sampling of the 2011 cohort has also increased the percentage of the sample from the South and the West and has increased the number of rural students at the expense of the urban students. Finally, the most dramatic change has been the increase in full day

Table 2. Unweighted Percentages of Demographic Characteristics of ECLS-K 1999 and ECLS-K 2011 Student Samples.

Characteristics	1999	2011
Gender		
Male	51.1	51.2
Female	48.9	48.8
Poverty		
Below poverty level	19.1	25.5
At or above poverty level	80.9	74.5
Age		
Up to 71 months	23.2	25.0
71–74 months	23.3	22.5
74–78 months	30.1	30.3
78 months and older	23.4	22.3
Race		
White	55.4	47.0
Black	15.2	13.3
Hispanic	17.8	25.0
Asian	7.5	9.2
American Indian or Native Hawaiian/Pacific Islander	1.8	1.0
Multiple race	2.4	4.4
Census region		
Northeast	18.4	16.7
Midwest	24.8	21.2
South	33.4	36.3
West	23.5	25.8
Urbanicity		
Urban	41.3	34.0
Suburban	35.4	36.2
Rural	23.3	29.8
Type of program		
Full day	56.4	83.8
Morning	26.2	9.4
Afternoon	17.4	6.8

Note. ECLS-K = early childhood longitudinal study kindergartners.

programs, with 84% of the 2011 cohort participating in full day programs compared to 56% in 1999.

Table 3 presents the descriptive statistics, including the means and *SDs*, of the fall and spring outcomes that are the focus of the analysis. The mean number of student observations for the spring measures was about 16,000, and the average number of schools was about 1,130. Across all outcomes,

Table 3. Descriptive Statistics of Fall and Spring Outcome Scores and Effect Size of Fall to Spring Change.

		Fall Pretest[b]		Spring Outcome		Fall to Spring Change[c]		Spring Sample	
Outcome	Range[a]	Mean	*SD*	Mean	*SD*	Effect Size[d]	*SE*	Schools[e]	Students[e]
Learning behaviors						**0.078**			
Approaches to learning (teacher report)	1.0–4.0	2.954	0.677	3.097	0.691	0.207	(.011)	1,054	15,978
Approaches to learning (parent report)	1.0–4.0	3.184	0.469	3.128	0.490	−0.115	(.012)	1,166	13,241
Attentional focus	1.0–7.0	4.732	1.314	4.920	1.332	0.141	(.011)	1,054	15,937
General behaviors						**0.123**			
Externalizing problem behaviors	1.0–4.0	1.593	0.622	1.638	0.641	0.069	(.011)	1,054	15,903
Inhibitory control	1.0–7.0	4.937	1.281	5.075	1.295	0.106	(.011)	1,052	15,925
Internalizing problem behaviors	1.0–4.0	1.458	0.488	1.510	0.497	0.104	(.011)	1,051	15,865
Interpersonal skills	1.0–4.0	3.001	0.635	3.133	0.654	0.202	(.011)	1,053	15,799
Self-control (teacher report)	1.0–4.0	3.091	0.624	3.178	0.637	0.137	(.011)	1,052	15,796
Self-control (parent report)	1.0–4.0	2.899	0.508	2.957	0.502	0.117	(.012)	1,166	13,254
Cognitive abilities						**0.205**			
Card sort postswitch score	0.0–6.0	5.246	1.656	5.555	1.205	0.256	(.013)	1,229	17,150
Card sort border score	0.0–6.0	3.725	1.200	4.088	1.317	0.276	(.011)	1,207	15,688
Numbers reversed score	40.0–175.0	93.542	16.612	94.983	17.139	0.084	(.011)	1,228	17,124
Academic abilities						**1.174**			
Math IRT score	5.3–74.9	29.416	10.799	42.026	11.116	1.134	(.012)	1,229	17,143
Reading IRT score	5.9–82.6	34.769	11.817	49.261	11.941	1.214	(.013)	1,229	17,185

Note. Effect size standard errors in parentheses. IRT = item response theory; *SD* = standard deviation; *SE* = standard error.
[a]Range is for both fall and spring Scores. [b]Fall statistics are only for those with spring data. [c]See text for calculation details. [d]Values in boldface indicates mean of outcome type. [e]Counts based on spring scores.

the *SDs* were similar between the fall and spring scores except for the cognitive measures.

Figure 1 presents the kernel density plots of the spring scores, student-level residuals, and school means for each of the 14 outcomes. While most of the outcomes follow an approximate normal distribution, several outcomes have skewed distributions. The most extreme example is the post-switch card score, where virtually all of the students scored six cards. The other scales that are skewed have means weighted toward the normative, for example, the modal approaches to learning scores (teacher rating) is the maximum, and the modal externalizing behavior score is the minimum. However, in most cases where the spring score was not normally distributed, the student residuals and school means did follow an approximate normal distribution. The one exception to this is again the card sort post-switch score, for which the school means are distributed as the spring outcome with most cases approaching the top score. However, it should be noted that the formal test of normality, the Shapiro–Wilks test, indicates that all outcomes, their student residuals, and school means, are not strictly normally distributed as all the z scores are larger than 2.

Outcome Gains in Effect Size Units

Table 3 also presents the differences between the spring and fall scores in effect size units (14). For each outcome type, the simple average of effect size is presented in boldface to offer a rough guide. Across the nonacademic outcomes, the outcome gains in effect size units range approximately between .1 and .2. The average for learning behaviors is lower because of the negative gains in the parent report of approaches to learning. General behaviors effect sizes have an average of about .12 and cognitive abilities increase by about 0.21 *SDs*.

The nonacademic effect sizes are much smaller than the academic effect sizes reported in this article and others (e.g., Bloom, Hill, Black, & Lipsey, 2008; Hill, Bloom, Black, & Lipsey, 2008). For the ECLS-K 2011 cohort, the Math and Reading IRT scales show large gains between fall and spring, as expected. The average effect size is greater than 1 but is slightly lower than other reported measures.

ICC and R^2 *Estimates*

Table 4 presents the estimates of the spring ICCs for the unconditional models and R^2 statistics for three different models (pretest only,

Table 4. Intraclass Correlations and R^2 Statistics for the ECLS-K 2011 Sample Spring Scores.

	Pretest Covariates[a]			Demographic Covariates[b]		Pretest and Demographic Covariates[a,b]	
Outcome	ICC	Student R^2	School R^2	Student R^2	School R^2	Student R^2	School R^2
Learning behaviors	**.062**	**.436**	**.578**	**.073**	**0.273**	**.446**	**.610**
Approaches to learning (teacher report)	.077 (.006)	.514 (.006)	.478 (.022)	.097 (.004)	.076 (.016)	.525 (.005)	.496 (.022)
Approaches to learning (parent report)	.044 (.005)	.320 (.007)	.700 (.015)	.035 (.003)	.616 (.018)	.326 (.007)	.773 (.012)
Attentional focus	.064 (.005)	.475 (.006)	.557 (.020)	.088 (.004)	.127 (.019)	.486 (.006)	.560 (.020)
General behaviors	**.080**	**.410**	**.610**	**.050**	**.115**	**.419**	**.630**
Externalizing problem behaviors	.075 (.006)	.503 (.006)	.692 (.016)	.067 (.004)	.117 (.019)	.515 (.006)	.717 (.015)
Inhibitory control	.067 (.005)	.499 (.006)	.523 (.021)	.092 (.004)	.073 (.015)	.513 (.006)	.517 (.021)
Internalizing problem behaviors	.097 (.007)	.282 (.006)	.646 (.018)	.012 (.002)	.021 (.009)	.289 (.006)	.662 (.017)
Interpersonal skills	.095 (.007)	.393 (.006)	.526 (.021)	.064 (.004)	.036 (.011)	.404 (.006)	.550 (.021)
Self-control (teacher report)	.107 (.007)	.396 (.006)	.544 (.021)	.054 (.004)	.070 (.015)	.404 (.006)	.547 (.021)
Self-control (parent report)	.036 (.005)	.388 (.007)	.730 (.014)	.013 (.002)	.371 (.022)	.390 (.007)	.788 (.011)
Cognitive abilities	**.071**	**.139**	**.611**	**.025**	**.578**	**.147**	**.719**
Card sort postswitch score	.048 (.005)	.060 (.004)	.577 (.018)	.006 (.001)	.290 (.022)	.060 (.004)	.549 (.019)
Card sort border score	.058 (.005)	.045 (.003)	.450 (.021)	.021 (.002)	.678 (.015)	.060 (.004)	.773 (.011)
Numbers reversed score	.106 (.007)	.311 (.006)	.805 (.010)	.049 (.003)	.765 (.012)	.320 (.006)	.835 (.009)
Academic abilities	**.192**	**.648**	**.658**	**.097**	**.507**	**.653**	**.679**
Math IRT score	.197 (.009)	.656 (.004)	.711 (.014)	.091 (.004)	.565 (.019)	.661 (.004)	.726 (.013)
Reading IRT score	.186 (.009)	.640 (.004)	.604 (.018)	.103 (.004)	.448 (.021)	.645 (.004)	.631 (.017)

Note. Standard errors in parentheses, boldface values indicate mean of outcome type. ECLS-K = early childhood longitudinal study kindergartners; ICC = intraclass correlation; IRT = item response theory.

[a]Pretest is the fall score. [b]Demographics included gender, socioeconomic status, and race.

demographics, and pretest plus demographics). The ICCs for math and reading are the largest (about .19), while the ICCs for most other nonacademic outcomes averaged about between .06 and .08. While small, there was some variability in the parameters. Although the teacher reported self-control had an ICC close to .11, the parent report was far lower (.04). A similar pattern was evident in the approaches to learning outcome, where the teacher report has a larger ICC than the parent report (.08 vs. .04).

The student-level effectiveness of pretests (labeled in the table as Student R^2) also varied across outcomes. The academic pretests were most effective (with values averaging .65). For learning and general behaviors, the R^2 averaged about .4. Using student demographics offered little effectiveness, but marginally increased the total effectiveness when combined with pretests.

The effectiveness of a school average of the pretest (labeled in the table as School R^2) indicates that using the school-average pretest in the evaluation analysis is a key strategy for reducing the impact of clustering on significance tests (Hedges & Hedberg, 2007; Hedges & Rhoads, 2010). For most outcomes, this parameter was high across both the academic, cognitive, and behavioral outcomes, with averages approximately .6. School means of demographic covariates were more effective at the school level, for cognitive abilities, compared to the student-level effectiveness. As with the student-level results, the most variation is explained in models that combine the demographic and pretest variables.

Implications of Design Parameters

The implications of these results are presented in Table 5, which contains the number of schools necessary to detect effect sizes of .1 and .2 for within-school samples of 15 or 30 students. These are plausible ranges as the average effect size (excluding parental report of approaches to learning, math and reading) is .15. The table is organized in much the same way as Table 4, showcasing the different models (without covariates, pretest, demographic, and pretest plus demographic covariates).[7]

Without covariates, the number of schools necessary to detect an effect size of .1 in nonacademic outcomes is about 400 for within-school samples of 15 and about 300 for within-school samples of 30. The sample size requirements for general behaviors are larger, about 450 for within-school samples of 15 and 350 for within-school samples of 30. The number of schools required to detect an effect size of .2 is smaller, with about 100 schools for within-school samples of 15 to detect an effect size of .1,

Table 5. Number of Schools Required to Detect 0.1 and 0.2 Effect Sizes Based on Estimated Design Parameters.

	Without Covariates		Pretest Covariates[a]		Demographic Covariates[b]		Pretest and Demographic Covariates[a,b]	
Outcome	ES = 0.1	ES = 0.2	ES = 0.1	ES = 0.2	ES = 0.1	ES = 0.2	ES = 0.1	ES = 0.2
$n = 15$ students per school								
Learning behaviors	**393**	**101**	**199**	**53**	**336**	**86**	**193**	**51**
Approaches to learning (TR)	438	112	224	58	400	102	216	56
Approaches to learning (PR)	342	88	180	48	250	64	170	46
Attentional focus	400	102	194	52	358	92	192	50
General behaviors	**445**	**113**	**218**	**57**	**415**	**106**	**212**	**56**
Externalizing problem behaviors	432	110	172	46	392	100	164	44
Inhibitory control	408	104	202	52	376	96	200	52
Internalizing problem behaviors	496	126	246	64	488	124	240	62
Interpersonal skills	490	124	260	68	468	120	250	66
Self-control (TR)	526	134	270	70	492	126	266	70
Self-control (PR)	318	82	156	42	274	70	150	40
Cognitive abilities	**419**	**107**	**247**	**65**	**274**	**71**	**225**	**59**
Card sort postswitch score	352	90	254	66	308	80	258	68
Card sort border score	382	98	292	76	254	66	230	60
Numbers reversed scale score	522	132	196	52	260	68	186	48
Academic abilities	**774**	**196**	**268**	**69**	**451**	**115**	**254**	**66**
Math IRT score	790	200	240	62	424	108	230	60
Reading IRT Score	758	192	296	76	478	122	278	72
$n = 30$ students per school								
Learning behaviors	**295**	**76**	**144**	**38**	**245**	**63**	**138**	**37**
Approaches to learning (TR)	342	88	176	46	314	80	170	46
Approaches to learning (PR)	242	62	112	30	152	40	102	28

(continued)

Table 5. (continued)

Outcome	Without Covariates		Pretest Covariates[a]		Demographic Covariates[b]		Pretest and Demographic Covariates[a,b]	
	ES = 0.1	ES = 0.2	ES = 0.1	ES = 0.2	ES = 0.1	ES = 0.2	ES = 0.1	ES = 0.2
Attentional focus	302	78	144	38	268	70	142	38
General behaviors	**349**	**89**	**161**	**42**	**323**	**83**	**156**	**42**
Externalizing problem behaviors	336	86	124	34	302	78	116	32
Inhibitory control	310	80	152	40	286	74	152	40
Internalizing problem behaviors	402	102	178	46	394	100	174	46
Interpersonal skills	396	102	202	52	380	98	194	52
Self-control (TR)	432	110	212	56	404	104	210	56
Self-control (PR)	216	56	96	26	174	46	88	26
Cognitive abilities	**323**	**83**	**163**	**43**	**179**	**47**	**141**	**39**
Card sort postswitch score	254	66	160	42	210	54	164	44
Card sort border score	284	74	198	52	158	42	138	38
Numbers reversed score	430	110	132	36	170	46	122	34
Academic abilities	**689**	**174**	**237**	**61**	**375**	**96**	**224**	**58**
Math IRT score	706	178	210	54	348	90	200	52
Reading IRT score	672	170	264	68	402	102	248	64

Note. Boldface numbers represent mean values by outcome types. Estimates are for 0.8 power for a two-tailed test with $\alpha = .05$. ES = effect size; IRT = item response theory.

[a]Pretest is the fall score. [b]Demographics included gender, socioeconomic status, and race.

and about 80 schools for within-school samples of 15 to detect an effect size of .2.

Including pretest and demographic covariates reduces the number of schools. The number of schools necessary to detect an effect size of .1 with covariates in nonacademic outcomes is about 200 for within-school samples of 15 and about 75 for within-school samples of 30. The number of schools required to detect an effect size of .2 with covariates is smaller, with about 50 schools for within-school samples of 15 to detect an effect size of .1 with covariates, and less than 50 schools for within-school samples of 15 to detect an effect size of .2 with covariates.

Changes to Math and Reading Parameters

The ICCs for both math and reading have decreased since the first ECLS-K cohort in 1999. In that survey, as reported in Hedges and Hedberg (2007), the ICC for math was .243 (standard error [*SE*] = .010) and it was 0.233 (*SE* = .010) for reading. In the current cohort, the ICC for math is .197 (*SE* = .009) and it is .186 (*SE* = .009) for reading. As reported in Table 6, these differences are beyond statistical chance, with approximate *z* statistics being 3.48 for math and 3.58 for reading (both statistically significant under a two-tailed test).

As reported in Table 2, the 1999 and 2011 samples are not equivalent. To explore whether the difference in the ICCs is due to some historical change in the school effects or due to the change in samples, conditional ICCs were calculated from six models each for math and reading. These models included one with the poverty indicator and its school mean, one with the race indicators and their school means, a model with census region indicators, urbanicity indicators, and type of program. These are the variables that changed the most from 1999 to 2011. Finally, a model that included all of the selected sample characteristics was estimated. For each model, the conditional ICC was calculated along with its *SE*. The poverty indicator model for math produced equivalent conditional ICCs for each cohort, and the full model produced equivalent ICCs for math. The differences in the conditional ICCs from the full model for reading were marginally significant, meaning that strictly speaking they were the same at conventional Type I error rates, but the difference is practically important. Thus, the tentative conclusion is that the change in sample may explain the differences for math score, but the answer is less conclusive for reading scores.

Table 6. Analysis of ICC Change for Academic Abilities from 1999 to 2011 by Select Covariates.

	Math				Reading			
Characteristics	1999	2011	z Test[b]	p[c]	1999	2011	z Test[b]	p[c]
Unconditional	.243 (.010)	.197 (.009)	3.479	.001	.233 (.010)	.186 (.009)	3.579	.000
Poverty indicator[a]	.147 (.008)	.134 (.008)	1.089	.276	.175 (.009)	.140 (.008)	2.927	.003
Race indicators[a]	.170 (.008)	.140 (.008)	2.682	.007	.204 (.009)	.146 (.008)	4.891	.000
Census region	.238 (.010)	.192 (.009)	3.515	.000	.232 (.009)	.185 (.009)	3.659	.000
Urbanicity	.237 (.010)	.192 (.009)	3.471	.001	.227 (.009)	.182 (.009)	3.586	.000
Type of program	.239 (.010)	.196 (.009)	3.204	.001	.229 (.010)	.184 (.009)	3.391	.001
Full model	.119 (.007)	.120 (.008)	−0.133	.894	.154 (.008)	.132 (.008)	1.897	.058

Note. Standard errors in parentheses. ICC = intraclass correlation.
[a]Schools means also included in the model. [b]z Test computed as difference divided by square root of the sum of the standard errors square. [c]p Values based on normal distribution, two-tailed test.

Discussion

The purpose of this article was to estimate the design parameters for behavioral and academic outcomes in early childhood, to explore the differences between academic and nonacademic outcomes, and finally to compare math and reading parameters between the 1999 and 2011 samples. This was done using the ECLS-K. The design of this study mimicked that of Hedges and Hedberg (2007) in that for each outcome, mixed models were fit to the outcomes to estimate ICCs and R^2 statistics. Effect size benchmarks were also estimated for fall–spring gains to provide upper bound estimates of sample sizes required for plausible effect sizes. These effect size benchmarks constitute a year's worth of growth and thus any intervention is likely to produce a smaller effect.

One expectation of the study was that school effects, and thus ICCs, would be minimal for behavioral outcomes. The ICCs were smaller for nonacademic outcomes than for math and reading, leading to the conclusion that schools correlate with these behaviors less than academic outcomes. This may be good news for planning research on behavioral outcomes, except for the result that the yearly gains in these outcomes is rather limited as measured by effect size. This means that the possible changes to behaviors (in the general population) are small, which can lead to larger required sample sizes despite the smaller ICCs.

This article provides the most systematic and representative catalogue of nonacademic design parameters to date. However, several nuances must be considered when utilizing these results. The first nuance is uncovered when we compare the teacher and parent reports of the same latent construct. The ICC for the teacher reported approaches to learning is .077, whereas the parent report is nearly half that at .044. The difference in self-control is even more striking, with the teacher report equaling .107 and the parent report is .036, nearly two thirds lower. It is difficult with the available data to know why these differences exist. One conservative possibility is that these ICCs simply reflect a teacher rater bias, that is, teachers tend to rate all their children as high or low on some scale. A more liberal interpretation is that children behave differently at home and school and that their school behaviors are influenced by peers and thus these ICCs are reflective of school conditions.

The math and reading parameters did change between cohorts, and this may be due to the differences in samples that resulted from changes to the demography and early childhood institutional organization between 1998 and 2011. A brief analysis to equate conditional ICCs was successful for

math, but less so for reading. Thus, further research on school effects and historical change must be explored. Finally, these parameters may not reflect the local parameters for a given study, but if our previous experience with math and reading are any indication, these values provide an upper bound for what researchers working in local contexts can expect. For example, Hedberg and Hedges (2014) found that within-district ICCs were lower that national estimates. As such, this article provides useful estimates for studying behavioral outcomes of young children using cluster randomized designs.

Appendix A

Review of Parameters Necessary to Plan Cluster Randomized Trials and Example Power Calculation

In cluster randomized trials, the appropriate analysis of the resulting data (without the use of covariates) is a multilevel model (also called a mixed or hierarchical model) where the outcome for the *i*th unit in cluster *j* is (using the notation of Raudenbush & Bryk, 2002)

$$y_{ij} = \gamma_{00} + \gamma_{01} T_j + u_{0j} + e_{ij},$$

where γ_{00} is the intercept term, γ_{01} is the treatment effect of the treatment indicator T_j at the cluster level, u_{0j} is the cluster-level random effect which is distributed normal with a mean of 0 and variance ${\sigma_2}^2$, and e_{ij} is the unit-level residual term with a mean of 0 and variance ${\sigma_1}^2$.

The statistical test of the treatment effect requires an estimate of the sampling variance of the treatment effect. The variance of the treatment effect (Raudenbush, 1997) employs the cluster- and unit-level variance, ${\sigma_2}^2$ and ${\sigma_1}^2$, respectively,

$$\text{var}\{\hat{\gamma}_{01}\} = \frac{2(\sigma_2^2 + \sigma_1^2/n)}{m},$$

where m is the number of clusters per treatment group and n is the number of units per cluster. The resulting t-test statistic (the effect estimate divided by the square root of the variance) is then

$$\hat{\gamma}_{01} \sqrt{\frac{m}{2(\hat{\sigma}_2^2 + \hat{\sigma}_1^2/n)}}$$

with $2m-2$ degrees of freedom (*df*). Unfortunately, to estimate the expected test statistic requires knowledge about the effect and the between- and within-cluster variances in the measurement units. A more useful approach to planning studies is to employ scale-free parameters.

The mean difference between treatment and control (i.e., the treatment effect), γ_{01}, can be expressed as scale-free effect size (using the total variation) via standardizing by the square root of the sum of the cluster-level variance, ${\sigma_2}^2$, and the unit-level variance, ${\sigma_1}^2$, (Hedges, 2007b)

$$\delta = \frac{\gamma_{01}}{\sqrt{\sigma_2^2 + \sigma_1^2}}.$$

The variance decomposition of the between- and within-cluster variance can also be made scale-free with the use of the intraclass correlation (ICC) parameter, which is the ratio of the between-cluster variance to the total variance

$$\rho = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2},$$

which facilitates the expression of the between-cluster variance, ${\sigma_2}^2$, as ρ and the within-cluster variance, ${\sigma_1}^2$, as $1 - \rho$. Given the above expressions, the scale-free expected test statistic (also the noncentrality parameter,λ) of the treatment effect can be expressed as (Hedges & Rhoads, 2010)

$$\lambda = \delta\sqrt{\frac{m}{2[\rho + (1-\rho)/n]}} = \delta\sqrt{\frac{mn}{2[1 + (n-1)\rho]}},$$

which follows a t distribution with $2m-2$ *df* and noncentrality parameter λ. This expression is more illustrative because it isolates the design effect, $1 + (n - 1)\rho$, which is a measure of how much the sampling variance increases due to cluster designs (Kish, 1974).

The covariates that are uncorrelated with treatment assignment are often employed to increase the precision of the treatment effect estimate without impacting the treatment effect estimate. One version of such an analysis is to choose unit-level variables,

x_{ij}, and their cluster means, $\bar{x}_j$, to enter into the model for the ith unit in cluster j

$$y_{ij} = \gamma_{00} + \gamma_{01} T_j + \sum_p \gamma_{p0} x_{pij} + \sum_p \gamma_{0p} \bar{x}_{pj} + u^*_{0j} + e^*_{ij},$$

where γ_{00} is the intercept term, γ_{01} is the treatment effect of the treatment indicator T_j at the cluster level, $\sum_p \gamma_{p0} x_{pij}$ is a set of p unit-level covariates and their associated effects, $\sum_p \gamma_{0p} \bar{x}_{pj}$ is a set of p cluster means of the unit covariates and their effects, u^*_{0j} is the cluster-level random effect which is distributed normal with a mean of 0 and variance $\tilde{\sigma}^2_2$, and e^*_{ij} is the unit-level residual term with a mean of 0 and variance $\tilde{\sigma}^2_1$.

The effectiveness of the covariates in reducing the between- and within-cluster variance can be expressed as the following R^2 statistics

$$R^2_2 = 1 - \frac{\tilde{\sigma}^2_2}{\sigma^2_2},$$

at the cluster level, and

$$R^2_1 = 1 - \frac{\tilde{\sigma}^2_1}{\sigma^2_1},$$

at unit level. With these R^2 statistics, the expected t-test statistic (also the noncentrality parameter) is (Hedges & Rhoads, 2010)

$$\lambda_A = \delta \sqrt{\frac{mn/2}{1 + (n-1)\rho - [R^2_1 + (nR^2_2 - R^2_1)\rho]}},$$

with $2m-2-p$ *df*. Note that the design effect, $1 + (n - 1)\rho$, is effectively adjusted in proportion to the effectiveness of the covariates by subtracting the quantity ${R_1}^2 + (n{R_2}^2 - {R_1}^2)\rho$.

With the noncentrality parameter in hand, an estimate of power for a two-tailed test for a given level of α (Type I error) and a given sample size of $2m$ clusters and n units within clusters (in the case of no covariates) can be accomplished with the cumulative distribution function of the noncentral t distribution

(H) with df and noncentrality parameter λ (Hedges & Rhoads, 2010)

$$p = 1 - H[t_{\frac{\alpha}{2},df}, df, \lambda] + H\left[-t_{\frac{\alpha}{2},df}, df, \lambda\right],$$

and power for a two-tailed test with covariates is

$$p = 1 - H\left[t_{\frac{\alpha}{2},df}, df, \lambda_A\right] + H\left[-t_{\frac{\alpha}{2},df}, df, \lambda_A\right],$$

where $t_{\frac{\alpha}{2},df}$ is the critical t value for a given level of α and df.

An example of a power calculation in *R* (Team, 2012/2014) for a design without covariates with $m = 8$, $n = 10$, an ICC of .2, and an effect size of .5 is

```
m <- 8 #set number of clusters per treatment
n <- 10 #set number of units per cluster
icc <- 0.2 #set icc
es <- 0.5 #set effect size
ncp <- es*sqrt(m*n/(2*(1+(n-1)*icc))) #calculate noncentrality
    parameter
df <- 2*m-2 #set the df
alpha <- 0.05 # set alpha
ct <- qt(alpha/2, df, lower.tail = FALSE) #get critical t value
power <- 1- pt(ct, df, ncp) + pt(-ct, df, ncp) #calculate power
power #show power
```

resulting in a power estimate of 0.42.

Appendix B

Missing Data Patterns for Outcomes by Demographic Characteristics

Table B1. Missing Patterns for Spring Outcome Scores by Demographic Characteristics.

		Gender		Type of Program			Race					
Outcome	Overall	Male	Female	Full day	Morning	Afternoon	White	Black	Hispanic	Asian	American Indian	Multiple race
Learning behaviors												
Approaches to learning (teacher report)	.12	.12	.12	.09	.10	.10	.09	.14	.14	.19	.17	.10
Approaches to learning (parent report)	.27	.27	.27	.28	.19	.23	.21	.39	.31	.30	.47	.17
Attentional focus	.12	.12	.12	.09	.11	.10	.09	.14	.15	.19	.18	.10
General behaviors												
Externalizing problem behaviors	.12	.13	.12	.09	.10	.10	.09	.14	.15	.19	.18	.11
Inhibitory control	.12	.13	.12	.09	.10	.10	.09	.14	.15	.19	.17	.11
Internalizing problem behaviors	.13	.13	.12	.10	.10	.10	.09	.15	.15	.20	.20	.11
Interpersonal skills	.13	.13	.13	.10	.10	.10	.10	.15	.16	0.20	.18	.11
Self-control (teacher report)	.13	.13	.13	.10	.11	.11	.10	.15	.16	.20	.18	.12

(continued)

Table B1. (continued)

Outcome	Overall	Gender		Type of Program			Race					
		Male	Female	Full day	Morning	Afternoon	White	Black	Hispanic	Asian	American Indian	Multiple race
Self-control (parent report)	.27	.26	.27	.28	.19	.23	.21	.39	.31	.30	.47	.17
Cognitive abilities												
Card sort postswitch score	.06	.06	.05	.05	.07	.06	.05	.07	.05	.08	.11	.05
Card sort border score	.14	.15	.12	.13	.14	.14	.11	.19	.15	.16	.17	.12
Numbers reversed score	.06	.06	.05	.05	.07	.06	.05	.07	.05	.08	.11	.05
Academic abilities												
Math IRT score	.06	.06	.05	.05	.07	.06	.05	.07	.05	.08	.11	.05
Reading IRT score	.05	.06	.05	.05	.06	.06	.05	.07	.05	.07	.11	.05

Outcome	Overall	Census region				Urbanicity			Age			
		Northeast	Midwest	South	West	Urban	Suburban	Rural	Up to 71 months	71–74 months	74–78 months	78 months and older
Learning behaviors												
Approaches to learning (teacher report)	.12	.13	.07	.09	.14	.14	.09	.07	.09	.08	.07	.07

(continued)

Table B1. (continued)

Outcome	Census region					Urbanicity			Age			
	Overall	Northeast	Midwest	South	West	Urban	Suburban	Rural	Up to 71 months	71–74 months	74–78 months	78 months and older
Learning behaviors												
Approaches to learning (parent report)	.27	.28	.24	.27	.26	.30	.24	.24	.26	.24	.26	.25
Attentional focus	.12	.13	.07	.09	.14	.15	.10	.07	.09	.08	.07	.07
General behaviors												
Externalizing problem behaviors	.12	.14	.07	.09	.14	.15	.09	.07	.10	.08	.08	.07
Inhibitory control	.12	.13	.07	.09	.14	.15	.10	.07	.09	.08	.08	.07
Internalizing problem behaviors	.13	.14	.07	.09	.14	.15	.10	.07	.10	.09	.08	.07
Interpersonal skills	.13	.14	.07	.10	.15	.16	.10	.08	.10	.09	.08	.08
Self-control (teacher report)	.13	.14	.07	.10	.15	.16	.10	.08	.10	.09	.08	.08
Self-control (parent report)	.27	.27	.24	0.27	.26	.30	.24	.24	.26	.24	.26	.25

(continued)

Table B1. (continued)

Outcome	Census region					Urbanicity			Age			
	Overall	Northeast	Midwest	South	West	Urban	Suburban	Rural	Up to 71 months	71–74 months	74–78 months	78 months and older
Cognitive abilities												
Card sort postswitch score	.06	.04	.05	.03	.04	.05	.03	.03	.00	.00	.00	.00
Card sort border score	.14	.14	.10	.11	.13	.14	.11	.10	.11	.08	.08	.08
Numbers reversed score	.06	.04	.05	.03	.04	.05	.04	.03	.01	.00	.00	.00
Academic abilities												
Math IRT score	.06	.04	.05	.03	.04	.05	.03	.03	.01	.00	.00	.01
Reading IRT score	.05	.03	.04	.03	.04	.05	.03	.03	.00	.00	.00	.00

Note. IRT = item response theory.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D140019, NORC at the University of Chicago. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute or the U.S. Department of Education.

Notes

1. Hedges and Hedberg (2007) provided evidence based on national probability samples for math and reading outcomes. Recent work has focused on deriving estimates based on state education agency administrative data sources for math and reading (Hedberg & Hedges, 2014; Hedges & Hedberg, 2013) and science (Westine, Spybrook, & Taylor, 2014) outcomes. Recently, this work has branched out from academic outcomes of children to other areas of education. For example, Kelcey and Phelps have estimated such parameters for use in professional development evaluations (2013a, 2013b).
2. The design of cluster randomized trials requires extant information about the variance decomposition of the outcome, typically operationalized as the intraclass correlation (ρ or ICC, and the effect of covariates on the outcome variance, operationalized as R^2 statistics, to estimate the power for a given sampling design (Raudenbush, 1997). Finally, a reasonable expectation about the effect size, δ, is also required. These parameters are defined in Appendix A.
3. The standard error (*SE*) of the effect size employs the sampling variance of the mean differences, $\text{var}\{\bar{y}_{\text{spring}} - \bar{y}_{\text{fall}}\} = N^{-1}(s^2_{\text{spring}} + s^2_{\text{fall}})$ and the sampling variance of the standard deviation, $\text{var}\{s_{\text{spring}}\} = [2(N-1)]^{-1}s_{\text{spring}}$ (Ahn & Fessler, 2003). Using the well-known delta method (see, e.g., Wolter, 2007, chapter 6), the variance of the effect size estimate is

$$\text{var}\{\text{ES}\} = (\text{ES})^2\left(\frac{\text{var}\{\bar{y}_{\text{spring}} - \bar{y}_{\text{fall}}\}}{(\bar{y}_{\text{spring}} - \bar{y}_{\text{fall}})^2} + \frac{\text{var}\{s_{\text{spring}}\}}{s^2_{\text{spring}}}\right),$$

   and its *SE* is the square root of the variance.
4. The variance of the estimated ICC employs the estimated sampling variance of the school-level variance component, $\text{var}\{\hat{\sigma}^2_2\}$ (see McCulloch, Searle, &

Neuhaus, 2008, for details on the sampling variance of the estimated variance of random effects) and is (Hedges, Hedberg, & Kuyper, 2012)

$$\text{var}\{\hat{\rho}\} = \frac{(1-\hat{\rho})^2 \text{var}\{\hat{\sigma}_2^2\}}{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)^2},$$

and its *SE* is the square root of this sampling variance.

5. The variances of the R^2 statistics are (see, e.g., Olkin & Finn, 1995)

$$\text{var}\{\hat{R}_1^2\} = \frac{4}{N}\hat{R}_1^2(1-\hat{R}_1^2)^2,$$

for the student level where N is the number of students, and

$$\text{var}\{\hat{R}_2^2\} = \frac{4}{J}\hat{R}_2^2(1-\hat{R}_2^2)^2,$$

for the school level where J is the total number of schools. The square root of these variances is the *SEs*.

6. For an example of how this impacts the ICCs, each of the released early childhood longitudinal study 1999 cohort data sets, the base-year file, the K–3 file, the K–5 file, and the K–8 file, were combined and the matched observations' item response theory scaled kindergarten scores from each file were analyzed using the kindergarten school (variable S2_ID). The ICCs decrease with each subsequent version of the rescaled score. Examining reading for cases in all data files, for instance, the base-year file's reading score's ICC is .220, the K–3 file's kindergarten ICC is .200, the K–5 file's kindergarten ICC is .194, and the K–8 file's kindergarten ICC is .186. At this time, the exact mechanism that leads rescaling to affect the variance components is not well documented, and further research is recommended.
7. Note that this table is most useful for nonacademic outcomes. If covariates are not available, the number of schools necessary to detect an effect size of 0.1 in math and reading is quite high (over 650 schools). However, given the large effect sizes for math and reading in Table 3, it is likely that an intervention's effect size would exceed 0.1.

References

Ahn, S., & Fessler, J. A. (2003). *Standard errors of mean, variance, and standard deviation estimators* (pp. 1–2). Ann Arbor: EECS Department, The University of Michigan.

Barrera, M. Jr., Biglan, A., Taylor, T. K., Gunn, B. K., Smolkowski, K., & Black, C., … Fowler, R. C. (2002). Early elementary school intervention to reduce conduct

problems: A randomized trial with Hispanic and non-Hispanic children. *Prevention Science*, *3*, 83–94.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness*, *1*, 289–328.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*, 415–427.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155.

Denham, S. A. (2006). Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early Education and Development*, *17*, 57–89.

Denham, S. A., & Brown, C. (2010). "Plays nice with others": Social–emotional learning and academic success. *Early Education and Development*, *21*, 652–680.

DiPerna, J. C., Lei, P.-W., & Reid, E. E. (2007). Kindergarten predictors of mathematical growth in the primary grades: An investigation using the Early Childhood Longitudinal Study—Kindergarten cohort. *Journal of Educational Psychology*, *99*, 369.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system: Manual*. Circle Pines, MN: American Guidance Service.

Hedberg, E., & Hedges, L. V. (2014). Reference values of within-district intraclass correlations of academic achievement by district characteristics results from a meta-analysis of district-specific values. *Evaluation Review*, *38*, 546–582.

Hedges, L. V. (2007a). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*, 151–179.

Hedges, L. V. (2007b). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60–87.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*, 445–489.

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of intraclass correlations in three-and four-level models. *Educational and Psychological Measurement, 72*, 893–909. doi:*0013164412445193*.

Hedges, L. V., & Rhoads, C. (2010). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness*, *3*, 157–198.

Kelcey, B., & Phelps, G. (2013a). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis*, *35*, 370–390.

Kelcey, B., & Phelps, G. (2013b). Strategies for improving power in school-randomized studies of professional development. *Evaluation Review*, *37*, 520–554.

Kish, L. (1974). *Survey sampling*. New York, NY: John Wiley.

McCulloch, C. E., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized linear, and mixed models*. New York, NY: Wiley-Interscience.

Mulligan, G. M., Hastedt, S., & McCarroll, J. C. (2012). *First-time kindergartners in 2010-11: First findings from the kindergarten rounds of the early childhood longitudinal study, kindergarten class of 2010-11* (ECLS-K: 2011; NCES 2012-049). Washington, DC: National Center for Education Statistics.

Najarian, M., Pollack, J. M., Sorongon, A. G., & Hausken, E. G. (2009). *Early childhood longitudinal study, kindergarten class of 1998–99 (ECLS-K): Psychometric report for the eighth grade*. Washington, DC: National Center for Education Statistics.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155.

Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the Children's Behavior Questionnaire. *Journal of Personality Assessment*, *87*, 102–112.

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*, 805–827.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173.

Raudenbush, S. W., & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Thousand Oaks, CA: Sage.

Rhoades, B. L., Greenberg, M. T., & Domitrovich, C. E. (2009). The contribution of inhibitory control to preschoolers' social–emotional competence. *Journal of Applied Developmental Psychology*, *30*, 310–320.

Rothbart, M. K., Ahadi, S. A., Hershey, K. L., & Fisher, P. (2001). Investigations of temperament at three to seven years: The Children's Behavior Questionnaire. *Child Development*, *72*, 1394–1408.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591–611.

Smith, A. L., Hoza, B., Linnea, K., McQuade, J. D., Tomb, M., & Vaughn, A. J., … Hook, H. (2013). Pilot physical activity intervention reduces severity of ADHD symptoms in young children. *Journal of Attention Disorders*, *17*, 70–82.

Stormont, M. (2002). Externalizing behavior problems in young children: Contributing factors and early intervention. *Psychology in the Schools*, *39*, 127–138.

Team, R. C. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://cran.r-project.Org (Original work published 2012)

Tourangeau, K., Nord, C., Lê, T., Sorongon, A., Hagedorn, M., Daly, P., & Najarian, M. (2012). *Early childhood longitudinal study, kindergarten class of 2010–11 (ECLSK: 2011), user's manual for the ECLS-K: 2011 kindergarten data file and electronic codebook (NCES 2013-061)*. Washington, DC: US Department of Education, National Center for Education Statistics.

Westine, C. D., Spybrook, J., & Taylor, J. A. (2014). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*. doi:10.1177/0193841X14531584.

What Works Clearinghouse. (2014). *WWC procedures and standards handbook*. Washington, DC: Author.

Wolter, K. (2007). *Introduction to variance estimation*. New York, NY: Springer Science & Business Media.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, *8*, 231–258.

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A method of assessing executive function in children. *Nature Protocols-Electronic Edition*, *1*, 297.

Author Biography

E. C. Hedberg is a senior research scientist at NORC at the University of Chicago. His research includes multilevel models, evaluation research, and social capital.